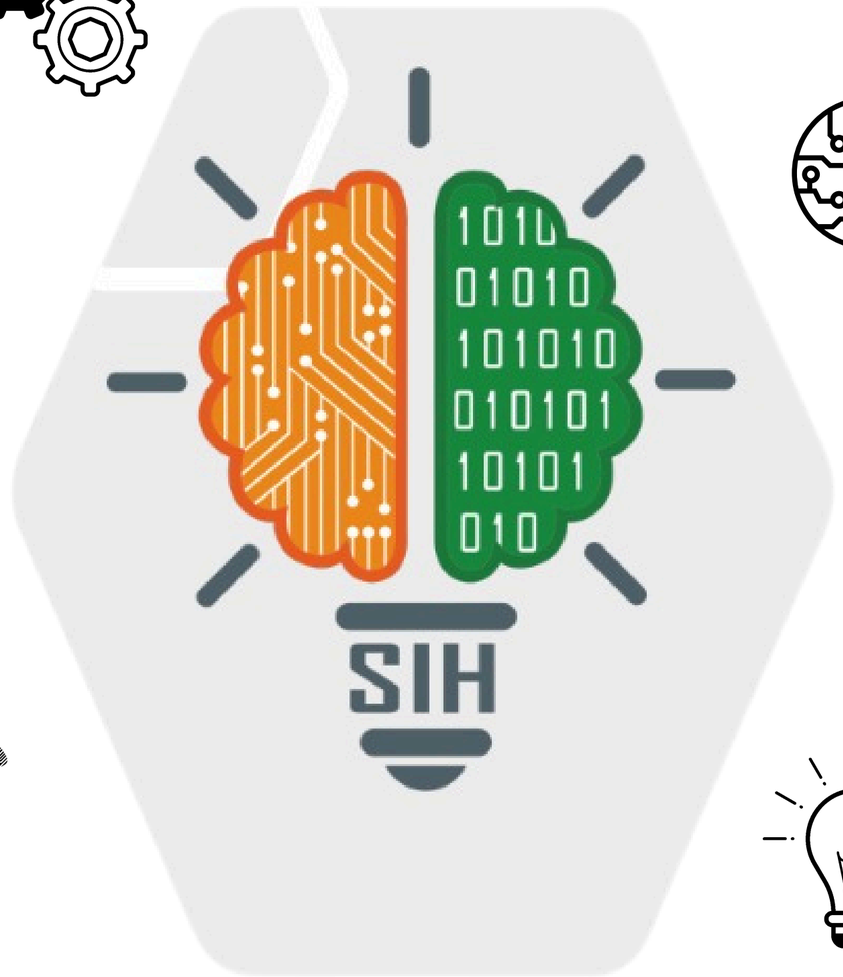
ALGOXEN

SMART INDIA HACKATHON 2024

SMART INDIA HACKATHON 2024

- **ORGANISATION** - Ministry Of Defence
- **PROBLEM STATEMENT TITLE** - Extraction and Verification of Information from Semi - Categorised data
- **PS NUMBER** - SIH1652
- **PS CATEGORY** - Software
- **THEME** - Smart Automation
- **TEAM NAME** - AlgoXen



ALGOXEN

SMART INDIA HACKATHON 2024

NSUT STUDENT PROFILE VERIFICATION MODEL

Objective: The model is designed to **verify** the authenticity and accuracy of information provided by students in their **profile application forms at NSUT, in place of the DRDO application form due to the limit of dataset available**.

Data Extraction

- Extracts key details such as the **student's name**, **date of birth**, and **gender** from the **NSUT Student Profile**

Cross-Verification

- Compares the extracted information with official documents like the **Aadhaar** card and **PAN** card.
- Ensures that the details (**name, date of birth, and gender**) match across all documents.It also check if photo on document is of same person or not

Consistency Check

- Verifies the uniformity of information across the **application form**, **Aadhaar**, and **PAN** cards to prevent discrepancies or errors.

Accuracy Assurance

- Helps maintain a reliable database by confirming that all **student profiles** are **accurate** and **consistent** with their official identification documents.

Enhanced Security

- **Reduces** the **risk** of fraudulent entries and ensures that the **student records** are **trustworthy** and **secure**..

Efficient Processing

- Automates the verification process to **save time** and **improve** the **efficiency** of the student registration system at **NSUT**.

ALGOXEN

SMART INDIA HACKATHON 2024

WORKFLOW (AI-ML MODEL)

INPUT

The model takes three inputs:
NSUT student profile Application form.
Aadhaar card image.
PAN card image.

TEXT EXTRACTION (Step 1)

Extract text from all three inputs using **Pytesseract.**

TEXT EXTRACTION (Step 2)

Use **SpaCy** to identify and extract specific entities such as the student's name, date of birth, and gender.

INFORMATION VERIFICATION (Step 3)

Compare the extracted name, date of birth, and gender from the application form with the corresponding details on the **Aadhaar** and **PAN** cards.

INFORMATION VERIFICATION (Step 4)

If the information **matches** across all documents, the model marks the **verification** as **successful**. If there are discrepancies, the model flags the m**ismatch** for **further reviews**.

ALGOXEN

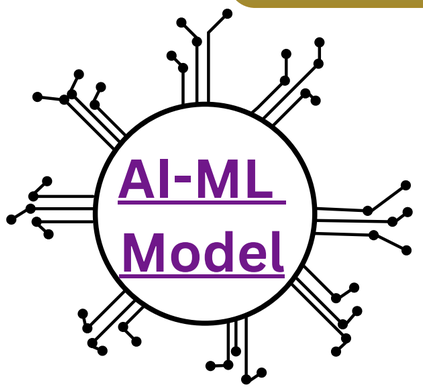
Technical Approach

SMART INDIA HACKATHON 2024

OCR
Optical character recognition

SpaCy
Utilised to identify and label specific entities such as name , gender , etc from the extracted text

NLP
Natural Language Processing which is used for identifying and labeling entities



DeepFace
enhances the accuracy of face matching by leveraging advanced facial recognition algorithms

Pytesseract
Extracting text from the application form , Aadhar card and Pan Card images

Label Studio
Labeling and categorising information to ensure correct identification and processing

NER
Named entity recognition

Front-End

The website features a sleek **React-based frontend** with intuitive ,user - friendly input fields for seamless document submission and verification. Built with **Tailwind CSS** for responsive design and styled components , it leverages efficient validation mechanisms.It also uses **javascript**

Back-End

The backend uses **Flask or Django** to connect with a machine learning model for document verification with **three sigma accuracy** . It communicates with the frontend via **RESTful API's** ensuring precise and reliable results.

ALGOXEN

TECH STACK

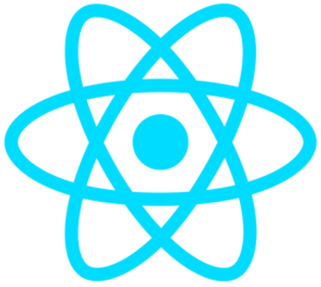
SMART INDIA HACKATHON 2024

Label Studio

Flask

spaCy

deepface



tailwindcss

ALGOXEN

FUTURE PROSPECTS

SMART INDIA HACKATHON 2024

Introduction to LayoutLM in the Student Profile Verification System

- In the **Student Profile Verification System**, **LayoutLM** is utilized to enhance the extraction of structured information from application forms. Unlike traditional Named Entity Recognition (**NER**) methods that focus solely on text, **LayoutLM** integrates both text and layout information from documents.
- This allows it to **accurately identify** and **extract** key fields such as **student name**, **date of birth**, and **gender**, which are often positioned in specific locations within forms.
- By leveraging **LayoutLM's** ability to understand spatial relationships and document structure, the system achieves **improved accuracy** in processing and verifying **student profiles** compared to conventional **NER** approaches.

(This is available free as a part of **HuggingFace** transformers library.)

IMPROVED ACCURACY

LayoutLM outperforms traditional **NER** methods by incorporating both text and layout, leading to **higher accuracy** in recognizing and extracting information, especially in cases where text positioning is crucial.

ENHANCED CONTEXTUAL UNDERSTANDING

LayoutLM enhances accuracy by combining text and layout data, allowing for more **precise extraction** of dates, names, and personal details.

SPATIAL AWARENESS

LayoutLM uses layout information to **accurately extract fields** from structured forms by understanding text positioning and organization.

ALGOXEN

SMART INDIA HACKATHON 2024

01 AZURE FORM RECOGNIZER

Automatically extracts text, key-value pairs, and table data from documents such as application forms, **Aadhaar** cards, and **PAN** cards, using pre-built models for efficient identity document processing and structured information extraction.

02 AZURE COGNITIVE SERVICES (TEXT ANALYTICS)

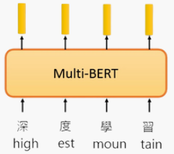
Provides Named Entity Recognition (**NER**) to identify and categorize entities such as names, dates, and gender directly from the extracted text.

03 AZURE COGNITIVE SERVICES (CUSTOM VISION)

If further customization is required, **Azure's Custom Vision service** can be trained to recognize specific forms and document types.

04 INDICBERT AND MBERT

These are **multilingual** models pre-trained on Indian languages. They can be fine-tuned for **NER** tasks to extract specific information from documents in Indian languages.



SUPPORTED LANGUAGES

IndicBERT is trained on **12** major Indian languages:
Hindi, **Bengali**, **Marathi**, **Tamil**, **Telugu**, **Gujarati**, **Kannada**, **Malayalam**, **Odia**, **Punjabi**, **Assamese** and **English** (as part of the multilingual corpus)

ALGOXEN

LONG-TERM IMPACTS AND BENEFITS OF KEY TECHNOLOGIES

SMART INDIA HACKATHON 2024

OCR & NLP (Pytesseract, SpaCy, LayoutLM)

- **Scalability & Adaptability**: Efficiently processes large volumes and adapts to new formats and languages.
- **Accuracy & Efficiency**: Automates verification, reducing errors, speeding up processes, and cutting costs.

Azure Cognitive Services

- **Future-Proofing**: Stays updated with AI advancements, ensuring long-term relevance.
- **Security & Compliance**: Robust cloud security for data protection and regulatory compliance.

Facial Recognition (DeepFace)

- **Fraud Reduction**: Minimizes identity fraud with advanced verification techniques.
- **Continuous Improvement:** AI algorithms enhance accuracy over time.

Frontend & Backend (React, Flask/Django)

- **User Experience**: Ensures a responsive and intuitive interface across devices.
- **Modular & Expandable**: Easy to update and add new features, ensuring long-term flexibility.

Document Classification (Azure Form Recognizer)

- **Automate Sorting**: Automatically classifies and organizes documents, improving workflow efficiency.
- **Enhanced Accuracy**: Reduces manual sorting errors, ensuring consistent and reliable results.

ALGOXEN

SMART INDIA HACKATHON 2024

OVERVIEW

- **The NSUT Profile Verification Model is a prototype for automated student document verification.**
- **It extracts information from an application form and verifies it against official government documents such as Aadhaar card, PAN card, and marksheets.**
- **With the right datasets, the model can be adapted to fully meet DRDO's requirements.**

BUSINESS PROSPECTS

Scalability
Handles large volumes, ideal for institutions of any size.

Adaptability
Can be tailored for sectors like finance, healthcare, and employment.

Accuracy & Security
Utilizes advanced OCR and NER for precise data extraction and robust security.

Versatility
Customizable for other institutions with appropriate datasets, broadening its application.

Future Potential
With additional resources, it can evolve into a comprehensive solution for DRDO and other sectors.

ALGOXEN

SMART INDIA HACKATHON 2024

Thank you very much!